Unconditionnally stable scheme
for Riccati equation

François Dubois [a] and Abdelkader Saïdi [b]

[a] *Conservatoire National des Arts et Métiers*

15 rue Marat, F-78 210 Saint Cyr l'Ecole, France.

[b] *Institut de Recherche Mathématique Avancée*

Université Louis Pasteur

7 Rue René-Descartes, 67084 Strasbourg Cedex, France.

July 2000 [*]

Abstract. – In this contribution we present a numerical scheme for the resolution of matrix Riccati equation used in control problems. The scheme is unconditionnally stable and the solution is definite positive at each time step of the resolution. We prove the convergence in the scalar case and present several numerical experiments for classical test cases.

Keywords: control problems, ordinary differential equations, stability.

AMS classification: 34H05, 49K15, 65L20, 93C15.

1)   Introduction

We study the optimal control of a differential linear system

$$(1) \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = A\,y + B\,v\,,$$

where the state variable $y(t)$ belongs to $\mathrm{I\!R}^n$ and the control function $v(\bullet)$ takes its values in $\mathrm{I\!R}^m$, with n and m beeing given integers. Matrix A is composed by n lines and n columns and matrix B contains n lines and m columns. Both matrices A and B are independent of time. With the ordinary differential equation (1) is associated an initial condition

$$(2) \qquad y(0) = y_0$$

with y_0 given in $\mathrm{I\!R}^n$ and the solution of system (1)(2) is parametrized by the function $v(\bullet)$: The control problem consists of finding the minimum $u(\bullet)$ of some quadratic functional $J(\bullet)$:

$$(3) \qquad J(u(\bullet)) \leq J(v(\bullet)), \quad \forall\, v(\bullet)\,.$$

The functional $J(\bullet)$ depends on the control variable function $v(\bullet)$, is defined by the horizon $T > 0$, the symmetric semi-definite positive n by n constant matrix Q and the symmetric definite positive m by m constant matrix R. We set classically :

$$(4) \qquad J(v(\bullet)) = \frac{1}{2}\int_0^T (Qy(t), y(t))\,\mathrm{d}t + \frac{1}{2}\int_0^T (Rv(t), v(t))\,\mathrm{d}t\,.$$

•   Problem (1)(2)(3)(4) is a classical linear quadratic mathematical modelling of dynamical systems in automatics (see *e.g.* Lewis [Le86]). When the control function $v(\bullet)$ is supposed to be square integrable $(v(\bullet) \in \mathrm{L}^2(]0; T[, \mathrm{I\!R}^m))$ then the control problem (1)(2)(3)(4) has a unique solution $u(\bullet) \in \mathrm{L}^2(]0; T[, \mathrm{I\!R}^m)$ (see for instance Lions [Li68]). When there is no constraint on the control variable the minimum $u(\bullet)$ of the functional $J(v)$ is characterized by the condition:

$$(5) \qquad \mathrm{d}J(u)\bullet w = 0\,, \quad \forall\, w \in L^2(]0, T[,\, \mathrm{I\!R}^m)\,,$$

which is not obvious to compute directly.

•   When we introduce the differential equation (1) as a constraint between $y(\bullet)$ and $v(\bullet)$, the associated Lagrange multiplyer $p(\bullet)$ is a function of time and is classically named the adjoint variable. Research of a minimum for $J(\bullet)$ (condition (5)) can be rewritten in the form of research of a saddle point and the evolution equation for the adjoint variable is classical (see *e.g.* Lewis [Le86]):

$$(6) \qquad \frac{\mathrm{d}p}{\mathrm{d}t} + A^{\mathrm{t}}p + Q\,y = 0\,,$$

with a final condition at $t = T$,

$$(7) \qquad p(T) = 0$$

and the optimal control in terms of the adjoint state $p(\bullet)$ takes the form:

$$(8) \qquad R\,u(t) + B^{\mathrm{t}}\,p(t) = 0\,.$$

• We observe that the differential system (1)(6) together with the initial condition (2) and the final condition (7) is coupled through the optimality condition (8). In practice, we need a linear feedback function of the state variable $y(t)$ instead of the adjoint variable $p(t)$. Because adjoint state $p(\bullet)$ depends linearly on state variable $y(\bullet)$ we can set:

$$p(t) \;=\; X(T-t) \bullet y(t) \,, \qquad 0 \le t \le T \,,$$

with a symmetric n by n matrix $X(\bullet)$ which is positive definite. The final condition (7) is realized for each value $y(T)$, then we have the following condition:

$$(9) \qquad\qquad\qquad\qquad X(0) \;=\; 0 \,.$$

We set $K = BR^{-1}B^{\mathrm{t}}$; we remark that matrix K is symmetric positive definite, we replace the control $u(t)$ by its value obtained in relation (8) and we deduce after elementary algebra the evolution equation for the transition matrix $X(\bullet)$:

$$(10) \qquad\qquad \frac{\mathrm{d}X}{\mathrm{d}t} \;-\; \left(XA + A^{\mathrm{t}}X \right) \;+\; XKX \;-\; Q \;=\; 0 \,,$$

which defines the Riccati equation associated with the control problem (1)(2)(3)(4).

• In this paper we study the numerical approximation of differential system (9)(10). Recall that the given matrices Q and K are $n \times n$ symmetric matrices, with Q semi-definite positive and K positive definite; the matrix A is an n by n matrix without any other condition and the unknown matrix $X(t)$ is symmetric. We have the following property (see *e.g.* Lewis [Le86]).

Proposition 1.   The solution of Riccati equation is positive definite.
Let K, Q, A be given $n \times n$ matrices with K, Q symmetric, Q positive and K definite positive. Let $X(\bullet)$ be the solution of the Riccati differential equation (10) with initial condition (9). Then $X(t)$ is well defined for any $t \ge 0$, is symmetric and for each $t > 0$, $X(t)$ is definite positive and tends to a definite positive matrix X_∞ as t tends to infinity: $X(t) \longrightarrow X_\infty$ if $t \longrightarrow \infty$. Matrix X_∞ is the unique positive symmetric matrix which is solution of the so-called algebraic Riccati equation:

$$-(XA + A^{\mathrm{t}}X) \;+\; XKX \;-\; Q \;=\; 0 \,.$$

• As a consequence of this proposition it is usefull to simplify the feedback command law (8) by the associated limit command obtained by taking $t \longrightarrow \infty$, that is:

$$(11) \qquad\qquad\qquad v(t) \;=\; -R^{-1}B^{\mathrm{t}} X_\infty \, y(t) \,,$$

and the differential system (1) (11) is stable (see *e.g.* [Le86]). The practical computation of matrix X_∞ by direct methods is not obvious and we refer *e.g.* to Laub [La79]. If we wish to compute directly a numerical solution of instationnary Riccati equation (10) classical methods for ordinary differential equations like *e.g.* the forward Euler method

$$\frac{1}{\Delta t}(X_{j+1} - X_j) \;+\; X_j K X_j \;-\; (A^t X_j + X_j A) \;-\; Q \;=\; 0 \,,$$

or Runge Kutta method fail to maintain positivity of the iterate X_{j+1} at the order $(j+1)$:

$$(12) \qquad\qquad (X_{j+1}\,x\,,\,x)\;>\;0\,,\quad \forall\,x\in\mathbb{R}^n,\quad x\neq 0\,,$$

if X_j is positive definite and if time step $\Delta t > 0$ is not small enough (see *e.g.* Dieci and Eirola [DE96]). Morever, there is to our best knowledge no simple way to determine *a priori* if time step $\Delta t > 0$ is compatible or not with condition (12).

• In the following, we propose a method for numerical integration of Riccati equation (10) which maintains condition (12) for each time step $\Delta t > 0$. We present in second section the simple case of scalar Riccati equation and present the numerical scheme and its principal properties of the general case in section 3. We describe several numerical experiments in section 4.

2)   Scalar Riccati equation

• When the unknown is a scalar variable, we write Riccati equation in the following form:

$$(13) \qquad\qquad \frac{\mathrm{d}x}{\mathrm{d}t}\,+\,k\,x^2\,-\,2\,a\,x\,-\,q\;=\;0\,,$$

with

$$(14) \qquad\qquad k > 0\,,\quad q \geq 0\,,$$

and an initial condition:

$$(15) \qquad\qquad x(0)\;=\;d,\quad d\geq 0\,.$$

We approach the ordinary differential equation (13) with a finite difference scheme of the type proposed by Baraille [Ba91] for hypersonic chemical kinetics and independently with the "family method" proposed by Cariolle [Ca79] and studied by Miellou [Mi84]. We suppose that time step Δt is strictly positive. The idea is to write the approximation x_{j+1} at time step $(j+1)\Delta t$ as a rational fraction of x_j with positive coefficients. We decompose first the real number a into positive and negative parts : $a = a^+ - a^-$; $a^+ = \max(0;a) \geq 0$, $a^- = \max(0;-a) \geq 0$, $a^+ a^- = 0$ and factorize the product x^2 into the very simple form:

$$\left(x^2\right)_{j+1/2}\;=\;x_j\,x_{j+1}\,.$$

Definition 1.   Numerical scheme in the scalar case.
For resolution of the scalar differential equation (13), we define our numerical scheme by the following relation:

$$(16) \qquad\quad \frac{x_{j+1}-x_j}{\Delta t}\,+\,k\,x_j\,x_{j+1}\,-\,2\,a^+\,x_j\,+\,2\,a^-\,x_{j+1}\,-\,q\;=\;0\,.$$

• The scheme (16) is implicit because some linear equation has to be solved to compute x_{j+1} when x_j is supposed to be given. In the case of our scheme this equation is linear and the solution x_{j+1} is obtained from scheme (16) by the homographic relation:

$$(17) \qquad\qquad x_{j+1}\;=\;\frac{\left(1\,+\,2a^+\,\Delta t\right)x_j\,+\,q\,\Delta t}{k\,\Delta t\,x_j\,+\,\left(1\,+\,2a^-\,\Delta t\right)}\,.$$

Proposition 2.   Algebraic properties of the scalar homographic scheme.
Let $(x_j)_{j \in \mathbb{N}}$ be the sequence defined by initial condition : $x_0 = x(0) = d$ and recurrence relation (17). Then sequence $(x_j)_{j \in \mathbb{N}}$ is globally defined and remains positive for each time step: $x_j > 0$, $\forall j \in \mathbb{N}$, $\forall \Delta t > 0$. If $\Delta t > 0$ is chosen such that:

$$(18) \qquad\qquad 1 + 2|a|\Delta t - k\, q\, \Delta t^2 \neq 0\,,$$

then $(x_j)_{j \in \mathbb{N}}$ converges towards the positive solution x^* of the "algebraic Riccati equation"

$$k\, x^2 - 2\, a\, x - q = 0$$

and

$$(19) \qquad\qquad x^* = \frac{1}{k}\left(a + \sqrt{a^2 + kq} \right).$$

- In the exceptional case where $\Delta t > 0$ is chosen such that (18) is not satisfied, then the sequence $(x_j)_{j \in \mathbb{N}}$ is equal to the constant $\frac{1 + 2\, a^+ \, \Delta t}{k\, \Delta t}$ for $j \geq 1$ and the scheme (17) cannot be used for the approximation of Riccati equation (13).

Theorem 1.   Convergence of the scalar scheme.
We suppose that the data k, a, q of Riccati equation satisfy (14) and (18) and that the datum d of condition (15) is relatively closed to x^*, *i.e.*:

$$(20) \qquad\qquad -\frac{1}{k\,\tau} + \eta \;\leq\; d - x^* \;\leq\; C\,,$$

where C is some given strictly positive constant $(C > 0)$, x^* calculated according to relation (19) is the limit in time of the Riccati equation, τ is defined from data k, a, q by:

$$\tau = \frac{1}{2\sqrt{a^2 + kq}}\,,$$

and η is some constant chosen such that

$$(21) \qquad\qquad 0 < \eta < \frac{1}{k\,\tau}\,.$$

- We denote by $x(t; d)$ the solution of differential equation (13) with initial condition (15). Let $(x_j(\Delta t\,; d_\Delta))_{(j \in \mathbb{N})}$ be the solution of the numerical scheme defined at the relation (17) and let d_Δ be the initial condition:

$$x_0(\Delta t\,; d_\Delta) = d_\Delta\,.$$

We suppose that the numerical initial condition $d_\Delta > 0$ satisfies a condition analogous to (20):

$$-\frac{1}{k\,\tau} + \eta \;\leq\; d_\Delta - x^* \;\leq\; C\,,$$

with C and $\eta > 0$ equal to the constant introduced in (20) and satisfying (21).

- Then the approximated value $(x_j(\Delta t\,; d_\Delta))_{j \in \mathbb{N}}$ is arbitrarily closed to the exact value $x(j\Delta t\,; d)$ for each j as $\Delta t \longrightarrow 0$ and $d_\Delta \longrightarrow d$. More precisely, if $a \neq 0$ we have the following estimate for the error at time equal to $j\Delta t$:

$$|x(j\Delta t\,; d) - x_j(\Delta t\,; d_\Delta)| \leq A\,(\Delta t + |d - d_\Delta|),\ \ \forall j \in \mathbb{N}, 0 < \Delta t \leq B$$

with constants $A > 0, B > 0$, depending on data k, a, q, η but independent on time step $\Delta t > 0$ and iteration j.

- If $a = 0$, the scheme is second order accurate in the following sense:

$$|x(j\Delta t; d) - x_j(\Delta t; d_\Delta)| \leq A\,(\Delta t^2 + |d - d_\Delta|), \ \ \forall\, j \in \mathbb{N}, 0 < \Delta t \leq B$$

with constants A et B independent on time step Δt and iteration j.

A direct application of the Lax theorem for numerical scheme associated to ordinary differential equations is not straightforward because both Riccati equation and the numerical scheme are nonlinear. Our proof is detailed in [DS2k].

3)   Matrix Riccati equation

In order to define a numerical scheme to solve the Riccati differential equation (10) with initial condition (9) we first introduce a strictly positive real number, which is chosen positive in such a way that the real matrix $[\mu\,\mathrm{I} - (A + A^\mathrm{t})]$ is definite positive:

$$(22) \qquad \frac{1}{2}(\mu\,x\,,\,x) - (A\,x\,,\,x) > 0\,, \quad \forall\,x \neq 0\,.$$

Then we introduce the definite positive matrix M wich depends on μ and matrix A:

$$M = \frac{1}{2}\mu\,\mathrm{I} - A\,.$$

The numerical scheme is then defined by analogy with relation (16). We have the following decomposition :

$$(23) \qquad A = A^+ - A^-$$

with $A^+ = \frac{1}{2}\mu\,\mathrm{I}$, $A^- = M$, $\mu > 0$, M positive definite. Taking as an explicit part the positive contribution A^+ of the decomposition (23) of matrix A and in the implicit part the negative contribution $A^- = M$ of the decomposition (23), we get

$$(24) \qquad \begin{cases} \dfrac{1}{\Delta t}(X_{j+1} - X_j) + \dfrac{1}{2}(X_j K X_{j+1} + X_{j+1} K X_j) + \\ \qquad\qquad\qquad + (M^\mathrm{t} X_{j+1} + X_{j+1} M) = \mu X_j + Q\,. \end{cases}$$

The numerical solution given by the scheme X_{j+1} at time step $j + 1$ is then defined as a solution of Lyapunov matrix equation with matrix X as unknown:

$$S_j^\mathrm{t}\,X + X\,S_j = Y_j$$

with

$$(25) \qquad S_j = \frac{1}{2}I + \frac{\Delta t}{2} K X_j + \Delta t\,M$$

and

$$(26) \qquad Y_j = X_j + \mu\,\Delta t\,X_j + \Delta t\,Q\,.$$

We notice that S_j is a (non necessarily symmetric) positive matrix and that Y_j is a symmetric definite positive matrix if it is the case for X_j.

Definition 2.   Symmetric matrices.
Let n be an integer greater or equal to 1. We define by $\mathcal{S}_n(\mathbb{R})$, (respectively $\mathcal{S}_n^+(\mathbb{R})$, $\mathcal{S}_n^{+*}(\mathbb{R})$) the linear space (respectively the closed cone, the open cone) of symmetric-matrices (respectively symmetric positive and symmetric definite positive matrices). The following inclusions $\mathcal{S}_n^{+*}(\mathbb{R}) \subset \mathcal{S}_n^+(\mathbb{R}) \subset \mathcal{S}_n(\mathbb{R})$ are natural.

Proposition 3.   Property of the Lyapunov equation.
Let S be a matrix which is not necessary symmetric, such that the associated quadratic form: $\mathbb{R}^n \ni x \longmapsto (x, Sx) \in \mathbb{R}$, is strictly positive *i.e.*

$$S + S^{\mathsf{t}} \in \mathcal{S}_n^{+*}(\mathbb{R}).$$

Then the application φ defined by :

(27) $$\mathcal{S}_n(\mathbb{R}) \ni X \longmapsto \varphi(X) = S^{\mathsf{t}} X + X S \in \mathcal{S}_n(\mathbb{R}),$$

is a one to one bijective application on the space $\mathcal{S}_n(\mathbb{R})$ of real symmetric matrices of order n. Morever, if matrix $\varphi(X)$ is positive (respectively definite positive) then the matrix X is also positive (respectively definite positive):

$$\text{if } \varphi(X) \in \mathcal{S}_n^+(\mathbb{R}), \quad \text{then } X \in \mathcal{S}_n^+(\mathbb{R}).$$

● The numerical scheme has been written as an equation with unknown $X = X_{j+1}$ which takes the form: $\varphi_j(X) = Y_j$ with φ_j given by a relation of the type (27) with the help of matrix S_j defined in (25) and a datum matrix Y_j defined by relation (26). Then we have the following propositions.

Proposition 4.   Homographic scheme computes a definite positive matrix.
The matrix X_j defined by numerical scheme (24) with the initial condition $X_0 = 0$ is positive for each time step $\Delta t > 0$:

$$X_j \in \mathcal{S}_n^+(\mathbb{R}), \qquad \forall\, j \geq 1.$$

If there exists some integer m such that X_m belongs to the open cone $\mathcal{S}_n^{+*}(\mathbb{R})$, then matrix X_{m+j} belongs to the open cone $\mathcal{S}_n^{+*}(\mathbb{R})$ for each j.

Proposition 5.   Monotonicity.
Under the condition $$\frac{1}{2}\left(K X_\infty + X_\infty K\right) < \left(\mu + \frac{1}{\Delta t}\right) I\,,$$

the scheme (24) is monotone and we have more precisely :

(28) $$\left(0 \leq X_j \leq X_\infty\right) \Longrightarrow \left(0 \leq X_j \leq X_{j+1} \leq X_\infty\right).$$

4)   First numerical experiments
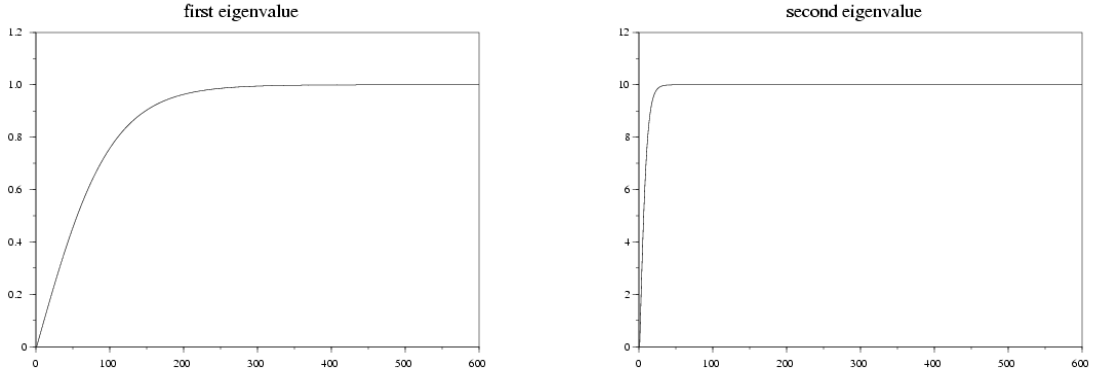
4-1   Square root function

● The first example studied is the resolution of the equation :

(29)
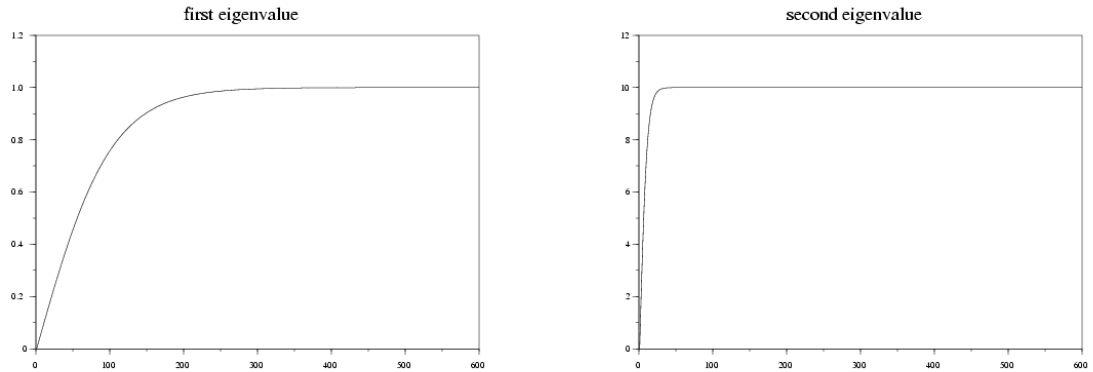$$\frac{\mathrm{d}X}{\mathrm{d}t} + X^2 - Q = 0, \quad X(0) = 0$$

with $n = 2$, $A = 0$, $K = I$ and matrix Q equal to

(30)
$$Q = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 100 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$
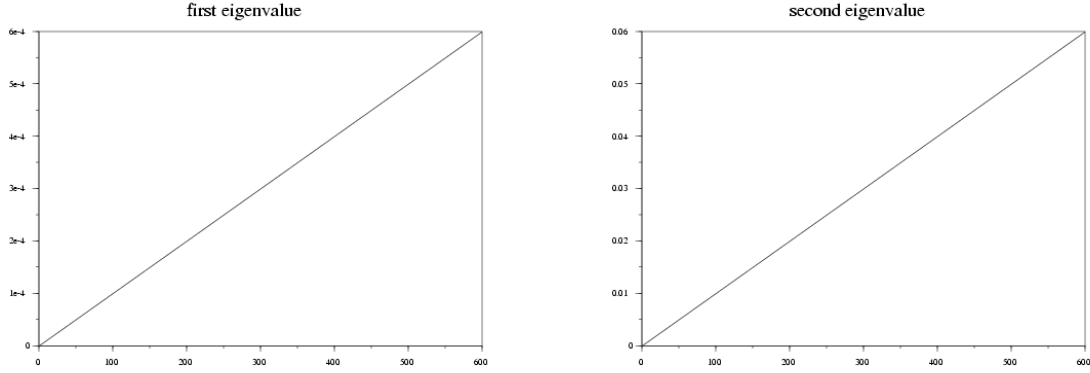
● We have tested our numerical scheme for fixed value $\Delta t = 1/100$ and different values of parameter μ : $\mu = 0.1$, 10^{-6}, 10^{+6}. For small values of parameter μ, the behaviour of the scheme does not change between $\mu = 0.1$ and $\mu = 10^{-6}$. Figures 1 to 4 show the evolution with time of the eigenvalues of matrix X_j and the convergence is achieved to the square root of matrix Q. For large value of parameter μ ($\mu = 10^{+6}$), we loose completely consistency of the scheme (see figures 5 and 6).

Figures 1 and 2.   Square root function test.
Two first eigenvalues of numerical solution ($\mu = 0.1$).

Figures 3 and 4.   Square root function test.
Two first eigenvalues of numerical solution ($\mu = 10^{-6}$).

Figures 5 and 6. Square root function test.
Two first eigenvalues of numerical solution ($\mu = 10^{+6}$).

4-2 Harmonic oscillator

● The second exemple is the classical harmonic oscillator. Dynamical system $y(t)$ is governed by the second order differential equation with command $v(t)$:

$$(31) \qquad \frac{\mathrm{d}^2 y(t)}{\mathrm{d}t^2} + 2\,\delta\,\frac{\mathrm{d}y(t)}{\mathrm{d}t} + \omega^2\,y(t) = b\,v(t)\,.$$

This equation is written as a first order system of differential equations :

$$(32) \qquad Y = \begin{pmatrix} y(t) \\ \frac{\mathrm{d}y(t)}{\mathrm{d}t} \end{pmatrix}, \quad \frac{\mathrm{d}Y}{\mathrm{d}t} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -2\,\delta \end{pmatrix} Y(t) + \begin{pmatrix} 0 \\ b\,v(t) \end{pmatrix}.$$

In this case, we have tested the stability of the scheme for fixed value of parameter $\mu\,(\mu = 0.1)$ and different values of time step Δt and coefficients of matrix R inside the cost function of relation (4):
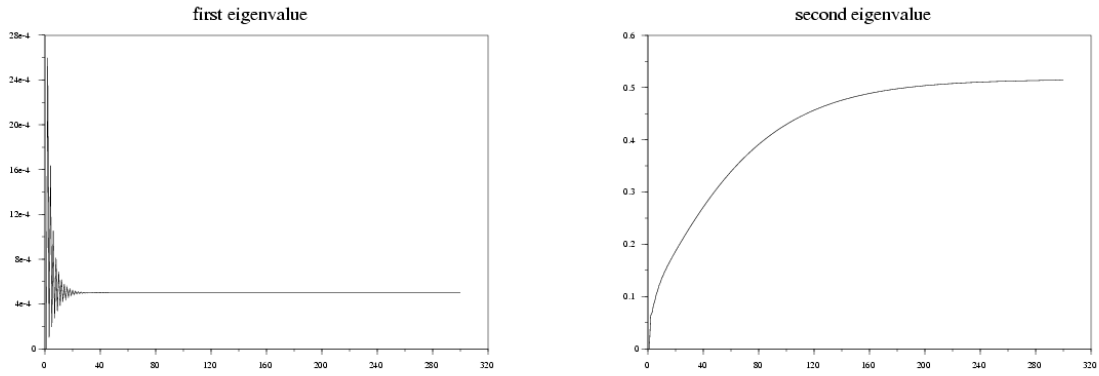
$$R = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}.$$

● We have chosen three sets of parameters : $\alpha = \Delta t = 1/100$ (reference experiment, figures 7 and 8), $\alpha = 10^{-6}$, $\Delta t = 1/100$ (very small value for α, figures 9 and 10) and $\alpha = 1/100$, $\Delta t = 100$ (too large value for time step, figures 11 and 12). Note that for the last set of parameters, classical explicit schemes fail to give any answer. As in previous test case, we have represented the two eigenvalues of discrete matrix solution X_j as time is increasing. On reference experiment (figures 7 and 8), we have convergence of the solution to the solution of algebraic Riccati equation. If control parameter α is chosen too small, the first eigenvalue of Riccati matrix oscillates during the first time steps but reach finally the correct values of limit matrix, the solution of algebraic Riccati equation. If time step is too large, we still have stability but we loose also monotonicity. Nevertheless, convergence is achieved as in previous case.
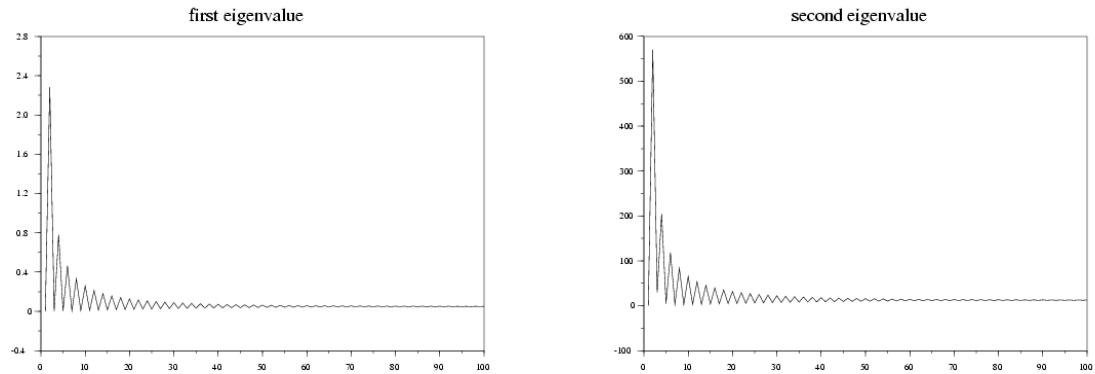
Figures 7 and 8.   Harmonic oscillator.

Two first eigenvalues of numerical solution ($\mu = 0.1$, $\alpha = 0.01$, $\Delta t = 0.01$).

Figures 9 and 10.   Harmonic oscillator.

Two first eigenvalues of numerical solution ($\mu = 0.1$, $\alpha = 10^{-6}$, $\Delta t = 0.01$).

Figures 11 and 12.   Harmonic oscillator.

Two first eigenvalues of numerical solution ($\mu = 0.1$, $\alpha = 0.01$, $\Delta t = 100$).

5)   Conclusion

We have proposed a numerical scheme for the resolution of the matrix Riccati equation. The scheme is implicit, unconditionnaly stable, needs to use only one scalar parameter and to solve a linear system of equations for each time step. This scheme is convergent in the scalar case and has good monotonicity properties in the matrix case. Our first numerical experiments show stability and robustness when various parameters have large variations. Situations where classical explicit schemes fail to give a solution compatible with the property that solution of Riccati equation is a definite positive matrix have been computed. We expect to prove convergence in the matrix case and we will present in [DS2k] experiments on realistic test models such as a string of vehicles and the discretized wave equation.

Acknowledgments

References

[Ba91] R. Baraille. Développement de schémas numériques adaptés à l'hydrodynamique, *Thèse de l'Université Bordeaux 1*, décembre 1991.

[Ca79] D. Cariolle. Modèle unidimentionnel de chimie de l'ozone, Internal note, *Etablissement d'Etudes et de Recherches Météorologiques*, Paris 1979.

[DE96] L. Dieci, T. Eirola. Preserving monotonicity in the numerical solution of Riccati differential equations, *Numer. Math.,* vol. 74, p. 35-47, 1996.

[DS2k] F. Dubois, A. Saïdi. Homographic scheme for Riccati equation, *CNAM-IAT Internal Report* number 338-2K, 29 august 2000, and *IRMA Research Report* number 2000-32, Université Louis Pasteur, Strasbourg, september 2000, hal-00554484, 2011.

[La79] A.J. Laub. A Schur Method for Solving Algebraic Riccati Equations, *IEEE Trans. Aut. Control*, vol. AC-24, p. 913-921, 1979.

[Le86] F.L. Lewis. *Optimal Control*, J. Wiley-Interscience, New York, 1986.

[Li68] J.L. Lions. *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[Mi84] J.C. Miellou. Existence globale pour une classe de systèmes paraboliques semi-linéaires modélisant le problème de la stratosphère : la méthode de la fonction agrégée, *C.R. Acad. Sci., Paris, Serie I*, t. 299, p.723-726, 1984.